Mittapalle, Kiran; Yagnavajjula, Madhu; Alku, Paavo

End-to-end Pathological Speech Detection using Wavelet Scattering Network

Published in:
IEEE Signal Processing Letters

DOI:
10.1109/LSP.2022.3199669

Published: 17/08/2022

Document Version
Publisher's PDF, also known as Version of record

End-to-End Pathological Speech Detection Using Wavelet Scattering Network

Mittapalle Kiran Reddy, Yagnavajjula Madhu Keerthana, and Paavo Alku, *Fellow, IEEE*

***Abstract*—In recent years, developing robust systems for automatic detection of pathological speech has attracted increasing interest among researchers and clinicians. This study proposes an end-to-end approach based on wavelet scattering network (WSN) for detection of pathological speech. In the proposed approach, the WSN (which involves no learning) extracts suitable information from the input raw speech signal and this information is then passed through a multi-layer perceptron (MLP) in order to classify the speech signal as either healthy or pathological. The results show that the proposed approach outperformed a convolutional neural network (CNN) based end-to-end system in distinguishing pathological speech from healthy speech. Furthermore, the proposed system achieved comparable performance with a state-of-the-art traditional system based on hand-crafted features for uncompressed speech, but gave better performance than the traditional system for compressed speech of low bit rates.**

***Index Terms*—Wavelet scattering network, CNN, pathological speech, MFCC, openSMILE features, MP3 compression.**

I. Introduction

SPEECH disorders are caused due to a disruption in the individual's ability to produce speech sounds precisely. Common conditions that can lead to speech disorders include developmental delays, heart failure, brain injuries, Parkinson's disease (PD), autism and laryngeal cancer [1], [2], [3], [4]. Speech disorders affect communication abilities of millions of people. Therefore, developing automatic methods for identifying people with speech disorders is essential to provide timely treatment, which can reduce symptoms and consequently help individuals to improve their speech communication.

In the literature, several methods have been proposed for automatic detection of pathological speech. These methods can be divided into two categories: traditional pipeline systems and modern end-to-end systems [4], [5]. The traditional pipeline systems consist of two stages. In the first stage, a set of hand-crafted features (like Mel-frequency cepstral coefficients (MFCCs)) are extracted from speech [3], [5], [15], [16], [17], [18], [19], [20]. In the second stage, a machine learning (ML) classifier, such as support vector machine (SVM), is trained with the selected features to label the input speech as either healthy or pathological [3], [4], [12], [13], [14]. A detailed review of various techniques considered for the feature extraction and ML classification parts of the traditional pipeline approach is given in [21].

In an end-to-end system, which is the focus of this study, the use of hand-crafted features is replaced by training deep learning models to directly map the raw speech signal (or the spectrogram) to the output labels (healthy/pathological). In order to develop deep learning models, existing studies have mainly used combinations of convolutional neural network (CNN) and multi-layer perceptron (MLP) [5], [6], [7]. The end-to-end systems are completely data-driven and they do not require any domain expertise in speech pathologies. However, unlike in many speech technology areas such as speech recognition, end-to-end systems have not been widely used in pathological speech detection. This is due to data scarcity, which is an inherent problem of pathological speech research since data is collected from patients whose condition might be so weak that recording large volumes of speech may not be possible [4], [5], [6].

As an alternative to CNN, we propose to use the wavelet scattering network (WSN) for end-to-end pathological speech detection. WSN is a deep convolution network, formed by a cascade of the wavelet transform (convolutions), modulus (non-linearity) and averaging (pooling) [8], [9]. It is capable of generating robust feature representations that are time-shift invariant and stable against time-warping deformations [8], [9]. WSN has a layer structure that is similar to CNN. However, WSN consists of fixed filters [8] and therefore unlike CNN it does not need training to extract features. This makes WSN an effective network to extract features for classification tasks in data-constrained scenarios. WSN has been previously utilized in tasks such as image classification [10] and electrocardiogram (ECG) beat classification [11]. To the best of our knowledge, this is the first study analyzing the effectiveness of WSN in pathological speech detection. The rest of the paper is organized as follows. The details of WSN and the proposed detection approach are described in Section II, followed by a discussion of experimental setup and results in Section III. Finally, Section IV concludes the current work and provides directions for future works.

II. Proposed End-to-End Pathological Speech Detection System Based on WSN

The proposed approach is based on capturing discriminative information from raw speech signals using WSN, followed by training an MLP with this information to predict one of the

Manuscript received 8 June 2022; revised 8 August 2022; accepted 13 August 2022. Date of publication 17 August 2022; date of current version 5 September 2022. The study was supported in part by the Academy of Finland under Project 330139, and in part by Aalto University (the MEC Program for India). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Corresponding author: Mittapalle Kiran Reddy.)*

Mittapalle Kiran Reddy and Paavo Alku are with the Department of Signal Processing and Acoustics, Aalto University, 00076 Aalto, Finland (e-mail: kiran.r.mittapalle@aalto.fi; paavo.alku@aalto.fi).

Yagnavajjula Madhu Keerthana is with the Department of Signal Processing and Acoustics, Aalto University, 00076 Aalto, Finland, and also with the Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India (e-mail: madhu.yagnavajjula@aalto.fi).

Digital Object Identifier 10.1109/LSP.2022.3199669

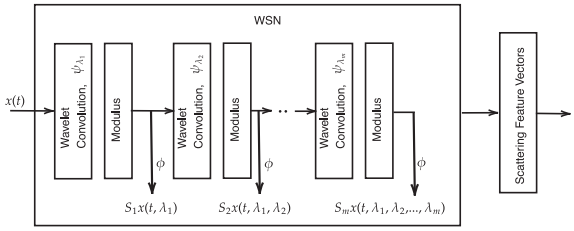

Fig. 1. A WSN iterates on wavelet modulus operators to generate scattering coefficient (feature) vectors for the input speech signal x.

two labels (healthy/pathological). These two main parts will be discussed in detail in the following two sub-sections.

A. Wsn

The WSN [8], [9] is a deep representation that processes data in layers (see Fig. 1), where each layer consists of three operations, namely, the wavelet transform, modulus, and averaging (with a low pass filter ϕ). The wavelet transform of input speech signal $x(t)$ is computed by convolving it with wavelets ψ_{λ_1} for $\lambda_1 \in \Lambda_1$, where Λ denotes the grid of all wavelet center frequencies λ_1 [8], [9]. This can be represented as

$$W_1 x = (x \star \psi_{\lambda_1}(t))_{t\in\mathbb{R},\lambda_1\in\Lambda_1} \tag{1}$$

Here, t is time index and wavelets ψ_{λ_1} have octave frequency resolution (or number of wavelet bandpass filters per octave) of Q_1. The wavelets used in this paper are the Gabor (analytic Morlet) wavelets. Next, the modulus of analytic wavelet coefficients is computed, and finally the previous calculation result is convolved with a low-pass filter ϕ (of length T and frequency bandwidth of $\frac{2\pi}{T}$), to produce the first-order scattering coefficients

$$S_1 x(t,\lambda_1) = |x \star \psi_{\lambda_1}(t)| \star \phi. \tag{2}$$

The time averaging process (shown in equation (2)) imposes time-shift in-variance and stability against time-warping deformations [8]. Convolving $|x \star \psi_{\lambda_1}(t)|$ with ϕ removes high-frequency components in order to localize the bins covering the lower spectrum [8]. The loss of high-frequency information in the first layer is recovered in the second layer by repeating the same operations as in the first layer but with $|x \star \psi_{\lambda_1}(t)|$ as input instead of $x(t)$. This generates second-order scattering coefficients as

$$S_2 x(t,\lambda_1,\lambda_2) = ||x \star \psi_{\lambda_1}(t)| \star \psi_{\lambda_2}(t)| \star \phi. \tag{3}$$

The wavelets ψ_{λ_2} have an octave frequency resolution Q_2 which may be different from Q_1. The second-order coefficients capture complementary high-frequency information, but there may be a loss of information in the second layer as well because of averaging. The lost information can be recovered by computing higher-order scattering coefficients. Higher orders proceed in the same manner by convolving wavelets with the modulus of the wavelet coefficients from the previous layer (see Fig. 1), followed by modulus and averaging. For any order $m \geq 1$, the scattering coefficients of order m can be obtained by

$$S_m x(t,\lambda_1,\ldots,\lambda_m) = ||\ldots|x \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_m}(t)| \star \phi(t). \tag{4}$$

Although scattering coefficients can be computed at any order, we use WSN with two layers in the current study. For the two-layer WSN, we set $Q_1 = 8$ and $Q_2 = 1$ for the wavelet filterbanks in the first (ψ_{λ_1}) and second layers (ψ_{λ_2}), respectively. The structure of the two filterbanks is shown in Fig. 2. The scattering cascade is similar to several neuro-physiological models of auditory processing, which incorporate cascades of constant-Q filterbanks followed by non-linearities [20], [21]. The first filterbank with $Q_1 = 8$ models the cochlear filtering and nearly corresponds to a mel-scale frequency subdivision [22], [23]. The second filterbank with $Q_1 = 1$ corresponds to later processing in the neuro-physiological models, which helps in characterizing the transients and attacks [8]. Improper vocal fold vibrations embed distinctive transient structures in pathological speech signals [5], [31], which can be represented effectively using the second filterbank. The two-layer WSN generates zeroth-, first- and second-order scattering coefficients for input speech signal x. Note that the zeroth-order coefficients are generated by a WSN through averaging of x with filter ϕ. However, in signals having little energy at low frequencies (like audio signals), these coefficients will be approximately zero [8] and hence they are not considered in this study.

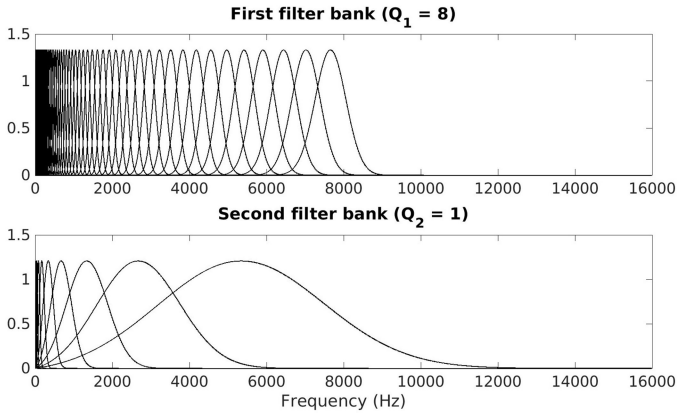

Fig. 2. Frequency responses of the first filterbank (top) and the second filterbank (bottom) of the two-layer WSN.

The number of scattering coefficients generated by a WSN is typically much larger compared to the length of x. Therefore, in practise, the network is critically down-sampled in time to provide a reduced representation that leads to a faster implementation [8], [9] without losing information. The amount of down-sampling depends on the bandwidth of ϕ, which relies on T [8], [9]. The larger the value of T, the fewer the number of scattering coefficients. But with a very large (or very small) T, the WSN do not capture enough information required for classification [8], [9]. In this work, we set $T = 250$ msec based on the experiments reported in Section III-D.

B. MLP for Detection

The final scattering representation of a speech signal is the aggregation of the first- and second- order scattering coefficients as an $N \times P$ feature matrix, where P is the number of time windows which varies with the length of the input signal and N denotes the fixed number of scattering coefficients in each time window. The considered two-layer WSN generates 260 scattering coefficients per time window (i.e. $N = 260$), for which logarithm is applied. For the detection task, a multi-layer perceptron (MLP) consisting of a single hidden layer with 256

hidden units is trained to predict one of the two labels (0: healthy or 1: pathological) for each time window. The MLP uses a rectified linear unit (ReLU) activation function for the hidden layer, and a softmax function for the output layer. The gradient tolerance and iteration limit for training were set to 10^{-6} and 10^{3}, respectively. The weights of the network were initialized with the Glorot initializer. The cross entropy loss was selected as the loss function, which was minimized using the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) [25]. During testing, the MLP first classifies each scattering time window of a new utterance separately. Then, majority voting is followed for obtaining the final binary decision (healthy/pathological) for the utterance.

III. Experiments and Results

A. Database

Experiments were conducted using the Saarbruecken Voice Database (SVD) [24], which is a corpus containing 71 different voice disorders. SVD includes recordings from 687 healthy speakers and 1356 voice disorder patients, collected from various speak tasks. In this study, we used the recordings from two speak tasks: (i) pronunciation of a German sentence "Guten Morgen, wie geht es Ihnen" ("Good morning, how are you?"), and (ii) sustained pronunciation of three vowels (/a/, /i/, /u/) in a constant neutral pitch. The data was down-sampled from 50 kHz to 16 kHz. As there are more recordings from patients than from healthy speakers, we balanced the class sizes by randomly selecting a subset of patients.

B. Baseline Systems

For comparison, we considered the recent CNN-based end-to-end pathological speech detection system (shortly referred to as CNNnet) proposed in [5]. The system takes speech segments (of duration 250 msec) as input, which are subsequently passed through three convolutional layers alternating with max-pooling layers (with size 2 and stride 2). The resulting feature representations from CNNs are passed through an MLP (with 256 hidden units) to predict the output label (healthy/pathological). A majority voting is used to obtain the binary decision at utterance level. The combination of CNNs and MLP is jointly trained from scratch in a single framework using the same hyper-parameters as in [5]. For complete details about the system, the reader is referred to [5].

In addition, comparison is also made with the recent traditional pipeline system developed in [26]. The system uses an SVM classifier (with polynomial kernel of order 2) and the state-of-the-art OpenSMILE features extracted from every utterance. In this work, separate systems are created by considering two sets of features extracted with the openSMILE toolkit [28]: (i) the 88-dimensional extended Geneva minimalistic acoustic parameter set (eGeMAPS) [27], and (ii) the 1582-dimensional INTERSPEECH 2010 paralinguistic challenge feature set (IS2010) [29].

C. Evaluation Metrics

For evaluation, five fold cross validation is used so that the recordings corresponding to 80% and 20% of all the speakers were used as training and testing data in each fold, respectively. Every speaker was used only once for testing and the same speaker was not used in both training and testing. Five standard performance metrics, namely, recall, precision, F1-score, accuracy and Mathews correlation coefficient (MCC) are considered for evaluating the performance [5], [30]. For a good detection system, recall, precision, F1-score and accuracy should be high (ideally 100%). The MCC metric varies between -1 (worst classifier) and 1 (best classifier), and should be close to 1 for a good detection system. Evaluation metrics were saved in each fold, and eventually averaged over all folds. The proposed and baseline systems were developed and evaluated separately for each of the considered speaking tasks.

TABLE I
Average Accuracy (In %) for Different T Values (In Msec)

T	50	100	150	200	250	300	350
Accuracy	72.17	74.04	75.47	75.18	75.33	74.10	74.10

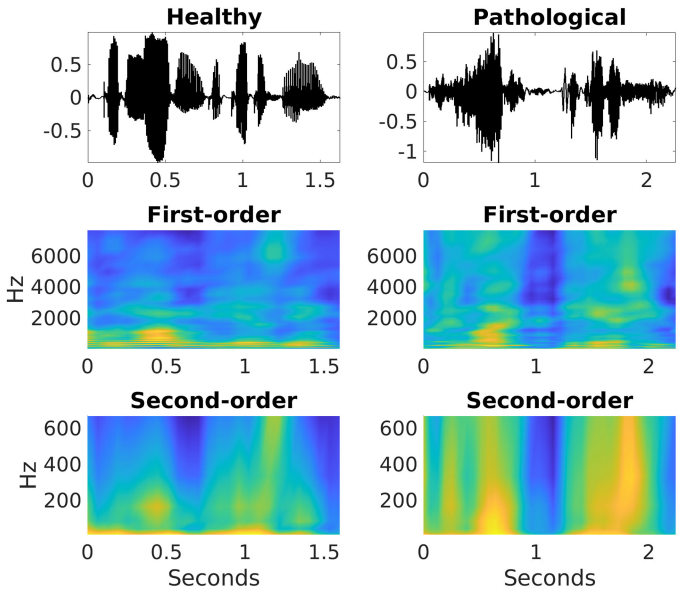

Fig. 3. Time-domain speech signals, the first- and second-order scattering coefficients corresponding to the German sentence "Guten Morgen, wie geht es Ihnen". The left and right panels show the plots for a healthy speaker and a patient with vocal fold paralysis having breathy voice, respectively. The scattering coefficients were computed using $Q_1 = 8$, $Q_2 = 1$ and $T = 250$ msec.

D. Results

First, we determined the optimal T parameter of WSN by computing the average cross-validation accuracies using T values varying from 50 to 350 msec in steps of 50 msec. This was done for a subset of SVD consisting of vowels from 600 randomly selected speakers (300 healthy and 300 patients). The results (shown in Table I) indicate that there is a negligible difference in accuracies for T values between 150–250 msec. We chose $T = 250$ msec, as it leads to generation of fewer scattering coefficients. Fig. 3 shows the plots of the first- and second- order scattering coefficients obtained from speech signals (sampled at 16 kHz) of a healthy individual and a patient uttering the same sentence using the chosen value of $T = 250$ msec. From the figure, it can be seen that the scattering coefficients provide an efficient representation and differentiation of the harmonic (shown by first-order coefficients) as well as transient structures

TABLE II
RESULTS OBTAINED WITH THE INDIVIDUAL DETECTION SYSTEMS. ALL THE METRICS ARE AVERAGED OVER THE FIVE FOLDS

	Recall	Precision	F1-score	Accuracy	MCC
Vowel task					
CNNnet	55.91	63.96	59.66 ± 2.5	62.09 ± 2.8	0.28
eGeMAPS	73.47	67.77	69.27 ± 2.9	70.49 ± 2.2	0.39
IS2010	77.07	70.42	73.59 ± 3.0	73.52 ± 2.7	0.53
Proposed	77.29	73.99	75.57 ± 1.5	75.01 ± 1.9	0.57
Sentence task					
CNNnet	55.56	72.16	62.78 ± 4.1	67.19 ± 2.7	0.32
eGeMAPS	77.07	70.42	73.59 ± 2.6	73.52 ± 2.4	0.47
IS2010	84.13	78.52	81.23 ± 1.3	80.63 ± 1.2	0.61
Proposed	81.46	81.25	81.34 ± 2.1	81.32 ± 2.2	0.63

(shown by second-order coefficients) of the healthy and pathological speech signals. This also justifies the use of a frame length of 250 msec in this study, since it is long enough to enable the WSN to effectively model the most discriminative information using a considerably fewer number of scattering coefficients.

Table II shows the detection results obtained for the individual detection systems for the SVD database. The proposed system provided better performance in terms of all the metrics compared to the baseline eGeMAPS system, for both the vowel and sentence tasks. Compared to the IS2010 system, the performance of the proposed system is moderately better for the vowel task and comparable for the sentence task. This indicates that the two-layer WSN yields robust features for classification with only a minimal set of user-specified parameters. In comparison with CNNnet, the proposed system provided $\approx 13\%$ higher detection accuracy for both speak tasks. The filter weights of deep CNNs can be learned well to extract discriminative features when the amount of training data is large, but large amounts of training data are typically not available for pathological speech. As a result, deep CNN-based systems under-perform in pathological speech detection. On the other hand, the WSN uses pre-defined wavelet filters (having a better physical meaning) as opposed to being learned as in standard CNNs. This enables extraction of highly discriminative features from raw speech segments, which can be easily visualized and interpreted, without any training data. For this reason, the proposed system has shown better detection results compared to CNNnet.

Nowadays, there is an increasing demand for remote patient monitoring (RPM), where data collected from patients is electronically transmitted to healthcare professionals. RPM applications have to deal with problems caused by compression (i.e. storing and transmitting speech). Therefore, we further investigate the proposed and IS2010 systems using compressed speech based on the MP3 compression scheme [32], [34]. The uncompressed speech signals sampled at 16 kHz from the sentence task of SVD were considered for experiments, with 70% of the data for training and the remaining 30% for testing. The systems were trained using uncompressed signals. The test speech signals were MP3-compressed using the Lame codec (version 3.100) [33] with an output sampling rate of 16 kHz and a resolution of 16 bits and using six different bit rates (128, 96, 64, 32, 16, and 8 kb/s). From Table III, it can be seen that the compression (down to at most 64 kbps) does not affect much the proposed system. This is because for the rates greater than or equal to 64 kbps there is no considerable loss of harmonic and transient information, as seen from Fig. 4. For the rates below 64 kbps, the information is progressively lost with a decrease

TABLE III
ACCURACY (%) OF THE PROPOSED AND BASELINE IS2010 SYSTEM FOR COMPRESSED SPEECH

Bit rate (kbps)	Compression ratio	Proposed	IS2010
8	32	67.23	62.38
16	16	70.39	65.53
32	8	74.51	71.36
64	4	78.89	77.67
96	2.66	79.37	78.64
128	2	79.61	79.13
Uncompressed	-	80.83	80.10

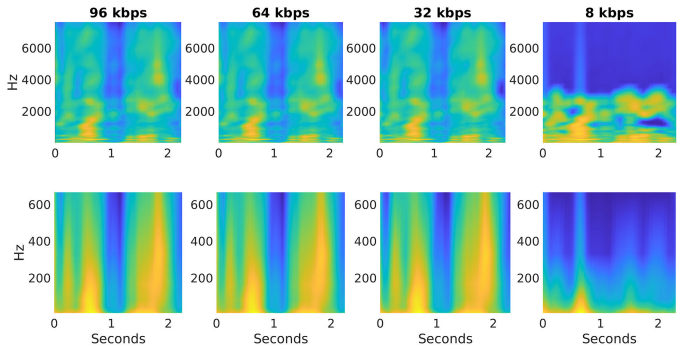

Fig. 4. Scattering representations of the MP3-compressed versions of the pathological speech signal shown in Fig. 3. The top and bottom rows show the plots for the first- and second-order coefficients, respectively.

in bit rate due to, for example, the reduction of the signal's audio bandwidth. Even at the bit rate of 8 kbps, the performance has not decreased drastically (seen from Table III) because the WSN still exhibits, to a fair extent, the harmonic and transient structures (see Fig. 4), which carry important discriminatory information for classification. The IS2010 system is also not much affected for bit rates $\geq$ 64 kbps and is comparable to the proposed system. However, for rates below 64 kbps, the proposed system achieves better accuracy than IS2010. This is because the features generated by a WSN are more stable than features such as MFCCs (included in the IS2010 feature set) to signal deformations [8], [9], which are notably present at low bit rates. Hence, the proposed system can be used effectively in tele-health applications.

IV. CONCLUSION

In this paper, we proposed a new end-to-end approach for pathological speech detection using a two-layer WSN, which can extract robust and discriminative speech feature representations and requires no training. The proposed system was compared with recent baseline pathological speech detection systems. The results show that the proposed end-to-end system outperformed the baseline CNN-based end-to-end system in discrimination of healthy and pathological speech signals. Furthermore, the performance of the proposed system is comparable to that of the IS2010 system for high bit rates (over 64 kbps), but it is better than the IS2010 system for very low bit rates. In the future, the performance of the proposed method may be examined for multi-class classification, and in prediction of severity of diseases such as PD and heart failure.

REFERENCES

[1] R. Chiaramonte and M. Bonfiglio, "Acoustic analysis of voice in parkinson's disease: A systematic review of voice disability and meta-analysis of studies," *Revue Neurologique*, vol. 70, no. 11, pp. 393–405, 2020.

[2] A. Blitzer et al., "Neurologic Disorders of the Larynx," New York, NY, USA: Thieme, 1992.

[3] M. K. Reddy et al., "The automatic detection of heart failure using speech signals," *Comput. Speech Lang.*, vol. 69, 2021, Art. no. 101205.

[4] M. K. Reddy and P. Alku, "A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation," *IEEE Access*, vol. 9, pp. 135953–135963, 2021.

[5] N. P. Narendra, B. Schuller, and P. Alku, "The detection of parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol 29, pp. 1925–1936, 2021.

[6] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 446–450.

[7] A. Rueda and S. Krishnan, "Augmenting dysphonia voice using fourier-based synchrosqueezing transform for a CNN classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6415–6419.

[8] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[9] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[10] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[11] Z. Liu et al., "Wavelet scattering transform for ECG beat classification," *Comput. Math. Methods Med.*, vol. 2020, 2020, Art. no. 3215681.

[12] L. A. Forero, M. Kohler, M. M. B. R. Vellasco, and E. Cataldo, "Analysis and classification of voice pathologies using glottal signal parameters," *J. Voice*, vol. 30, no. 5, pp. 549–556, 2016.

[13] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic detection of parkinson's disease based on modulated vowels," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 1190–1194.

[14] A. Mayle, Z. Mou, R. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 4514–4518.

[15] R. Orozco-Arroyave et al., "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.

[16] J. Mekyska et al., "Robust and complex approach of pathological speech signal analysis," *Neurocomputing*, vol. 167, pp. 94–111, 2015.

[17] R. Fraile et al., "Spectral analysis of pathological voices: Sustained vowels vs. running speech," in *Proc. Models Anal. Vocal Emissions Biomed. Appl.*, 2011, pp. 67–70 .

[18] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with parkinson's and neurological diseases using cepstral analysis," *IEEE Trans. Neural System Rehabiliation Eng.*, vol. 24, no. 10, pp. 1100–1108, Oct. 2016.

[19] J. I. Godino-Llorente and P. G. Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[20] I. A. Rezek and S. J. Roberts, "Stochastic complexity measures for physiological signal analysis," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 9, pp. 1186–1191, Sep. 1998.

[21] F. T. Al-Dhief et al., "A survey of voice pathology surveillance systems based on Internet of Things and machine learning algorithms," *IEEE Access*, vol. 8, pp. 64514–64533, 2020.

[22] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. detection and masking with narrow-band carriers," *J. Acoustical Soc. Amer.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[23] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[24] M. Pützer and W. J. Barry, "Saarbrücken voice database, institute of phonetics, university of saarland," Accessed: Dec. 2021. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/

[25] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., New York, NY, USA: Springer, 2006.

[26] P. Barche et al., "Towards automatic assessment of voice disorders: A clinical approach," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2537–2541.

[27] F. Eyben et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[28] "openSMILE 3.0.1"[online]. Available: https://github.com/audeering/opensmile/releases

[29] B. Schuller et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.

[30] S. Boughorbel et al., "Optimal classifier for imbalanced data using matthews correlation coefficient metric," *PLoS One*, vol. 12, no. 6, 2017, Art. no. e0177678.

[31] Z. Ali et al., "Voice pathology detection based on the modified voice contour and SVM," *Biologically Inspired Cogn. Architectures*, vol. 15, pp. 10–18, 2016.

[32] ISO-MPEG Audio Layer-3, *Information Technology-coding of Moving Pictures and Associated Audio for Digital Storage Media up to 1.5 Mbit/s. Part 3: Audio* ISO/IEC Standard 11172-3. Geneva, Switzerland: International Organization for Standardization, 1993.

[33] Lame MP3 Encoder, The lame project, 2017. [Online]. Available: https://lame.sourceforge.io/

[34] N. Saenz-Lechon et al., "Effects of audio compression in automatic detection of voice pathologies," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 12, pp. 2831–2835, Dec. 2008.